CLAIMS

What is claimed is:

1. A computer implemented method of merging with a base assembly of molecules one or more additional assemblies of molecules, similar molecules in the assemblies having previously been identified and removed using a validated molecular structural descriptor, comprising the steps of:

a. Using a validated molecular structural descriptor which is appropriate to whole molecules, characterizing all the molecules in the base assembly of molecules and in the assembly of molecules to be merged;

b. Calculating the molecular structural distance between every molecule in the base assembly to every molecule in the assembly to be merged;

c. While there are still molecules in the assembly to be merged which have not been tested, selecting a molecule from the assembly to be merged;

d. Determining whether the molecular structural distance between the selected molecule and every molecule in the base assembly is within the neighborhood distance of the molecular structural descriptor;

e. Select for inclusion in the merged assemblies only those molecules identified in step **d** as having molecular structural distances greater than the neighborhood distance.

f. Repeat step c through step e until all molecules in the assembly to be merged have been tested; and

g. Repeat step a through step f for each additional assembly to be merged.

2. The method of claim 1 in which the molecular structural descriptor appropriate to whole

molecules in the Tanimoto similarity coefficient.

3. A computer implemented method of merging with a base assembly of molecules one or more additional assemblies of molecules, similar molecules in one or more of the assemblies having not previously been identified and removed using a validated molecular structural descriptor, comprising the steps of:

a. Selecting subsets of each assembly by:

(1) Selecting a molecule within each assembly;

(2) Using a validated molecular structural descriptor appropriate to whole molecules, calculating the descriptor distance between the selected molecule and all molecules within the assembly;

(3) Determining the shortest distance between the selected molecule and all molecules previously selected for the subset;

(4) Selecting for inclusion in the subset the molecule whose shortest descriptor distance from the previously selected molecules is the largest and is greater than the neighborhood distance of the descriptor;

(5) Repeat steps (1) through (4) until the largest shortest difference between molecules is less than the neighborhood distance of the descriptor; and

(6) Repeat steps (1) through (5) for each assembly;

b. Using a validated molecular structural descriptor which is appropriate to whole molecules, characterizing all the molecules in the base assembly of molecules and in the assembly of molecules to be merged;

c. Calculating the molecular structural distance between every molecule in the base

assembly to every molecule in the assembly to be merged;

d. While there are still molecules in the assembly to be merged which have not been tested, selecting a molecule from the assembly to be merged;

e. Determining whether the molecular structural distance between the selected molecule and every molecule in the base assembly is within the neighborhood distance of the molecular structural descriptor;

f. Select for inclusion in the merged assemblies only those molecules identified in step e as having molecular structural distances greater than the neighborhood distance.

g. Repeat step d through step f until all molecules in the assembly to be merged have been tested; and

h. Repeat step b through step g for each additional assembly to be merged.